

Shuffled2026: a wrong-TCR decoy set and a Bayesian-network recognition classifier for TCR–pMHC complexes

tcren supplementary note

1 The Shuffled2026 decoy set

Real crystal structures give only positive examples of TCR recognition. To obtain physically-grounded *negatives* without any binding assay, we build **Shuffled2026** from the oriented Canonical2026 crystals (`tcren.shuffle` / `tcren shuffle`). Every complex is first superposed into the common canonical MHC frame; we then keep each complex’s **pMHC intact** (a real, correctly-presented peptide on its own MHC) and graft on a **different** complex’s TCR, drawn from a within-MHC-class derangement so no decoy reproduces a real pairing. Because all structures share the one canonical frame, the graft is a *direct chain replacement with no per-pair alignment* — unlike a pairwise MHC superposition, this lets each grafted TCR keep its own native MHC–TCR docking angle, so the decoy set spans the real docking-angle variance across the database. The peptide:MHC interface is untouched by construction (a built-in control: $F_{\text{pep:MHC}}$ and $\Delta F_{\text{pep:MHC}}$ are invariant), while both the TCR:peptide and TCR:MHC interfaces are mismatched. We generate 10× decoys per pMHC.

2 A conditional-linear-Gaussian Bayesian network

We classify real ($y = 1$) vs shuffled ($y = 0$) with a Gaussian Bayesian network over the full fresh `tcren` interface feature set (interface size and burial, dual-chain balance, the CDR-loop TCRen energies F and their poly-alanine ΔF for TCR:peptide / TCR:MHC / peptide:MHC, the TCRdock docking geometry, and the chemical contact-type counts) together with the **MHC-class** label. A DAG is learned by BIC hill-climbing on the within-class-centred, standardised features; the binary class y and the MHC class are then discrete parents of every feature node, shifting its conditional mean. Classification is the Gaussian log-likelihood ratio $\log P(x \mid y=1) - \log P(x \mid y=0)$. The network is shown in Figure 1; performance under 5-fold cross-validation is in Figure 2 and Table 1.

Table 1: Balanced cross-validated metrics (threshold at the Youden point for the point metrics).

metric	value
ROC–AUC	0.865
PR–AUC (baseline 0.091)	0.494
balanced accuracy	0.797
F_1	0.417
MCC	0.391

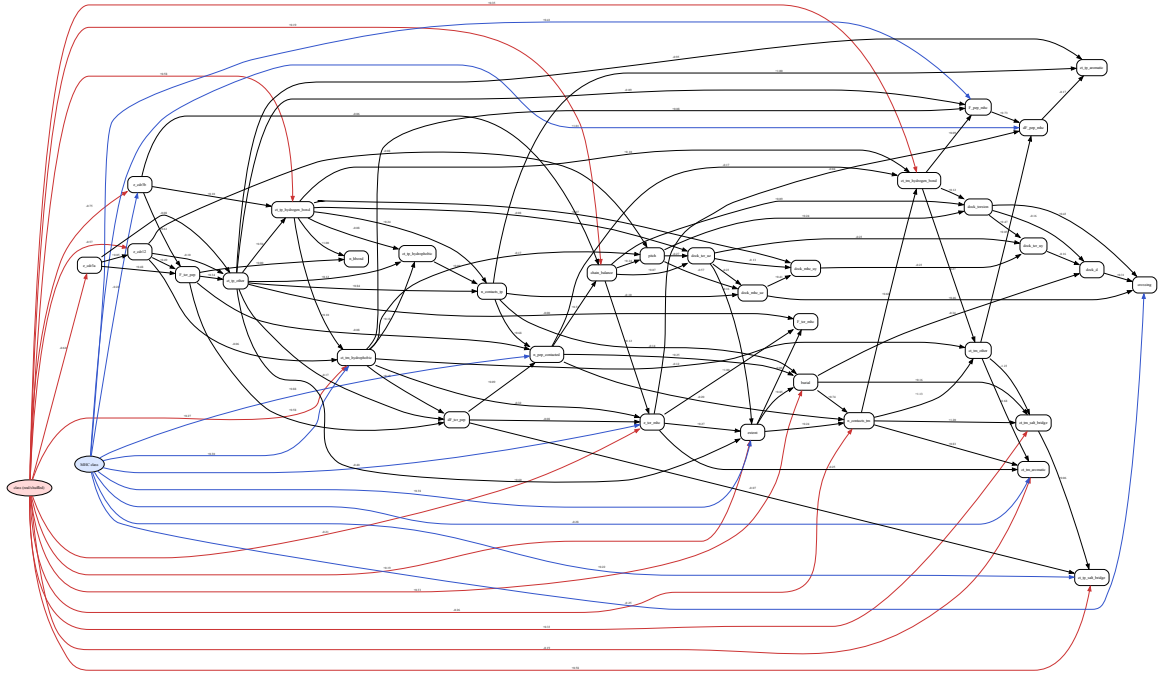
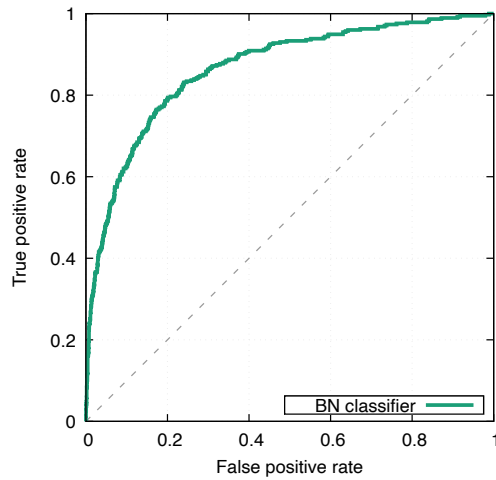
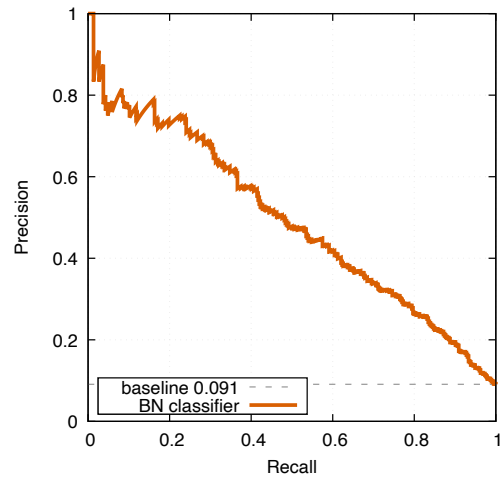


Figure 1: Learned Bayesian network. Boxes are interface features; the class node (red) and MHC-class node (blue) shift each feature's conditional mean (edge labels are standardised slopes; feature-feature edges are partial slopes).



(a) ROC



(b) Precision-recall

Figure 2: Real-vs-shuffled discrimination, 5-fold cross-validated out-of-fold predictions.

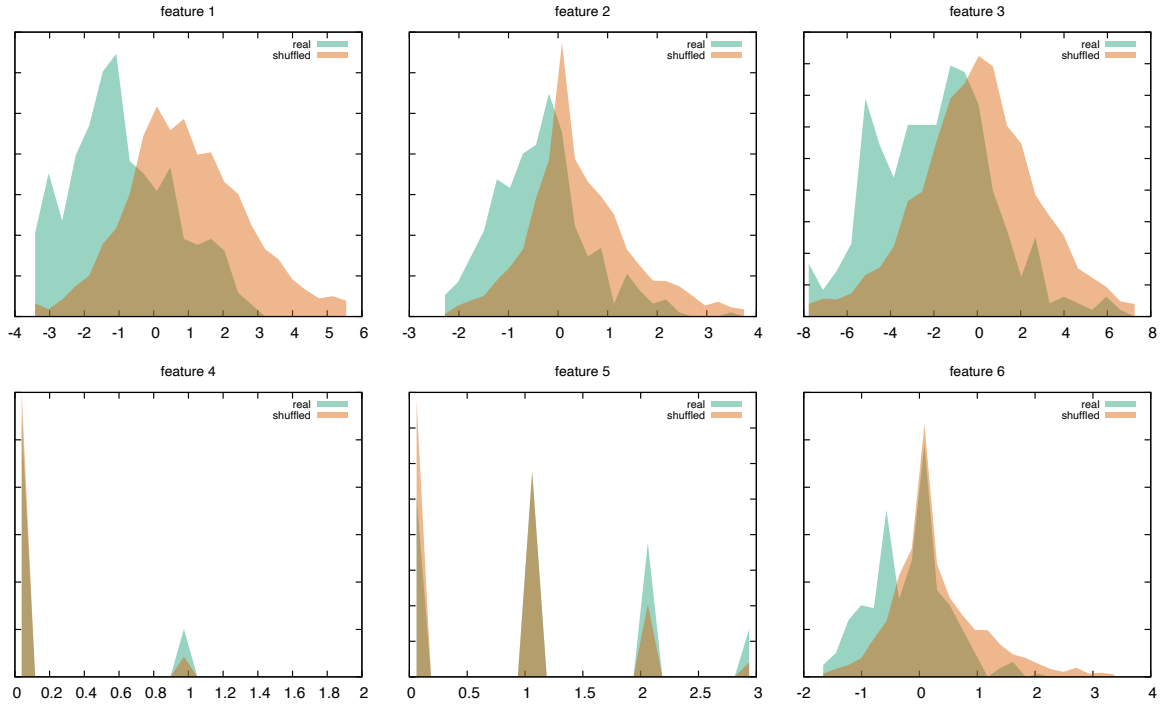


Figure 3: Marginal distributions of the six most class-separating features, real (green) vs shuffled (orange). Panels: (1) `F_tcr_pep`, (2) `e_cdr3b`, (3) `dF_tcr_pep`, (4) `ct_tp_salt_bridge`, (5) `ct_tm_salt_bridge`, (6) `e_cdr3a`.